1. Understanding Principal Components Analysis for spectral analysis

1.1. Preamble

The goal of this document is to understand Principal Components Analysis (PCA) in `scikit-learn.decomposition`.

I have done this by writing a sequence of tests that helped me understand what was going on. Take a look

This document is a work in progress and sections will improve as my understanding improves.

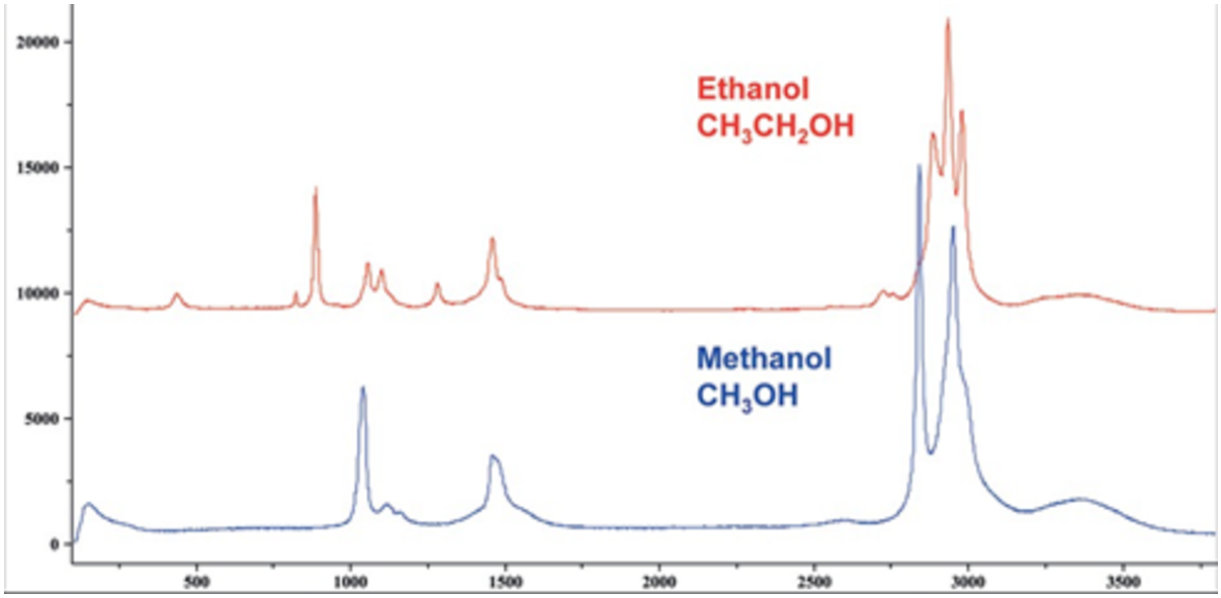
I worked on a presentaation of Dirac Notation that I will eventually use for this document, it is available here: https://www.icloud.com/keynote/0a7T8YWw1ZTV5_LU2YuABNHOQ#Notation_de_Dirac

1.2. Note

Use Typora to read this document to see the equations.

1.3. Introduction

Spectroscopy is the optical method of choice to detect substances or identify tissues. We have learned from very early on that identifying the peaks in a spectrum can be done to infer the substances in our sample. For instance, if we have pure substances, it is relatively easy to identify Ethanol and Methanol with their Raman spectra, because their shapes are significantly different and many peaks do not overlap:



It may even be possible to separate both substances if we have a mixture of the two, by fitting $c_e\mathbf{I}_(\nu)_e + c_m\mathbf{I}(\nu)_m$ to find the appropriate concentrations that can explain our final combined spectrum. However, what do we do if we have a mixture of several solutions? What if several peaks overlap? What if we don't know the original spectra that are being mixed?

We know intuitively that if peaks belong to the same molecule, they should vary together. If by chance none of the peaks from the different analytes overlap, then it becomes trivial: we only need to identify the peaks, find their amplitudes, and we will quickly get the concentrations of the respective analytes. But things get complicated if they have overlapping peaks, and even worse if we have more than a few components. We will discuss how we can use the very general nomenclature of a generalized vector space with Principal Components Analysis to extract this information.

1.4. Spectra as vectors

From a mathematical point of view, we can consider a spectrum as a **vector** of intensities:

$$\mathbf{I} = \sum_{i=1}^{n} I_i \hat{\nu}_i, \tag{1}$$

where each individual frequency ν_i is in its own dimension, with $\hat{\nu}_i$ the base vectors and I_i is the intensity at that frequency. Therefore, if we have N=1024 points in our intensity spectrum $\mathbf{I}$, we are in an N-dimensional space, with the components being $(I_1, I_2, \dots I_N)$, and we should assume (at least for now) that these components are all independent. If we define the **norm** of a vector from the **dot product**, we can say that the norm is equal to:

$$|\mathbf{I}| = \sqrt{\sum_{i=1}^{n} I_i \hat{\nu}_i \cdot \sum_{j=1}^{n} I_j \hat{\nu}_j} = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} I_i I_j \, \hat{\nu}_i \cdot \hat{\nu}_j} = \sqrt{\sum_{i=1}^{n} |I_i|^2}, \tag{2}$$

since the spectral base vectors $\hat{\nu}_i$ are all orthonormal, which we can use to normalize a spectrum (or a vector). Finally, it is very convenient to work with matrix notation to express many of these things. We can express the spectrum $\mathbf{I}$ in the basis $\{\hat{\nu}_i\}$ with:

$$\mathbf{I} = \sum_{i=1}^{n} I_i \hat{\nu}_i = \left(\hat{\nu}_1 \, \hat{\nu}_2 \dots \hat{\nu}_n\right)\left(I_1 \, I_2 \dots \, I_n\right)^T = \left(\hat{\nu}_1 \, \hat{\nu}_2 \dots \hat{\nu}_n\right) \begin{pmatrix} I_1 \\ I_2 \\ \dots \\ I_n \end{pmatrix}. \tag{3}$$

Note that the vector itself $\mathbf{I}$ is different from the *components of that vector in a given basis* $[I]_\nu$: the position *vector* $\mathbf{r}$ is different from its *components* (x, y, z) in Cartesian coordinates. If we consider these matrices as partitions, we can write in a form even more compact as:

$$\mathbf{I} = \hat{\nu}[I]_\nu \tag{4}$$

where the notation $[I]_\nu$ means "the intensity coefficients in base ν to multiply the base vectors $\hat{\nu}$ and obtain the vector (spectrum)". We will use the transpose notation to keep expressions on a single line when needed.

For more information about the notation for vector, base vectors and coefficients:

- Read Greenberg Section 10.7 on bases and base changes.
- Watch the video (in French) that explains in even more details where this comes from.
- Watch an example (in French) for problem 10.7.1 that discussed an application of this notation and formalism to perform a base change.

1.5. Bases

This document will always discuss things in terms of bases, base change: a vector can be expressed in an infinite number of bases. A reminder for the definition of a base $\{\mathbf{e}_i\}$:

1. A base set is **complete**: it spans the space for which it is a base: you must be able to get every vector in that space with $\mathbf{v} = \sum c_i \mathbf{e}_i$. We call the c_i the components of a vector *in that base*.
2. A base set is **linearly independent**: all base vectors are independent, and the only way to combine the base vectors to obtain the null vector $\sum c_i \mathbf{e}_i = \mathbf{0}$ is with $c_i = 0$ for all c_i. Another way of saying this is "you cannot combine two vectors from the base set to obtain a third one from the base set".
3. The number of base vectors in the set is the **dimension** of the space.

Notice that :

1. The base vectors **do not have to be unitary**: they can have any length. A **normalized** base set will be labelled with a hat on the vector $\left\{\hat{\mathbf{e}}_i\right\}$, and an arbitrary set will be $\{\mathbf{e}_i\}$.
2. The base vectors **do not have to be orthogonal**: as long as they are independent, that is fine. There is no notation to differentiate orthogonal and non-orthogonal basis because the property of orthogonality is not a

single vector property, it is a property of a pair of vectors, therefore we cannot label a vector as "orthogonal".

1.6. Spectra as *dependent* vectors

We know from experience that in a spectrum, intensities are not completely independent: for instance, in the methanol spectrum above, the peak around 1000 cm $^{-1}$ has a certain width and therefore those intensities are related and are not independent. In fact, for the spectrum of a single substance, *all intensities* are related because they will come from a scaled version of the original spectrum. Therefore, if we have the a methanol spectrum :

$$\mathbf{I}_M = \sum_{i=0}^{n} I_{M,i}\hat{\nu}_i, \tag{5}$$

where $I_{M,i}$ is the intensity at frequency ν_i. We can normalize this spectrum and obtain a "reference spectrum" or a "base spectrum" $\hat{\mathbf{s}}_M$ for a unity concentration :

$$\hat{\mathbf{s}}_M = \frac{\mathbf{I}_M}{|I_M|}. \tag{6}$$

Any other solution of methanol of scalar concentration c_M would simply yield the spectrum:

$$\mathbf{I} = c_M\hat{\mathbf{s}}_M = c_M\frac{\mathbf{I}_M}{|I_M|}. \tag{7}$$

So if we have several individual solutions with their spectra $\left\{\hat{\mathbf{s}}_j\right\}$ from which we create a mixture of concentrations c_j, we will generate spectra in a sub-space of the original n-dimensional intensity vector-space. The set of vectors $\left\{\hat{\mathbf{s}}_j\right\}$ is a basis set because we can generate all vectors in that sub-space with a linear combination of the base vectors (or base spectra). The dimension of that subspace is equal to the number of elements in $\left\{\hat{\mathbf{s}}_j\right\}$ We can write the mixture spectrum $\mathbf{I}$ as:

$$\mathbf{I} = \sum_j c_j \hat{\mathbf{s}}_j = \left(\hat{\mathbf{s}}_1, \hat{\mathbf{s}}_2, \ldots, \hat{\mathbf{s}}_n\right)(c_1, c_2, \ldots, c_n)^T = \hat{\mathbf{s}}[c]_{\hat{\mathbf{s}}} \tag{8}$$

Again, we read the last expression $[c]_{\hat{\mathbf{s}}}$ as "the coefficients in base $\left\{\hat{\mathbf{s}}\right\}$ needed to multiply the base vectors $\hat{\mathbf{s}}_i$ to obtain the final spectrum $\mathbf{I}$". It stresses the point that the vector $\mathbf{I}$ and its components in a given basis $[c]_{\hat{\mathbf{s}}}$ are not the same thing. This will become critical below when we look a Principal Components.

Finally, if we want to describe a **collection** of m spectra obtained from mixing these base solutions $\hat{\mathbf{s}}$ with concentrations c_{ij} for the i-th spectrum and the j-th base solution, we can write:

$$\mathbf{I}_i = \sum_j c_{ij} \hat{\mathbf{s}}_j. \tag{9}$$

This can be rewritten in matrix notation:

$$(\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_m) = \left(\hat{\mathbf{s}}_1, \hat{\mathbf{s}}_2, \ldots, \hat{\mathbf{s}}_n\right) \begin{pmatrix} c_{11} & c_{21} & \ldots & c_{m1} \\ c_{12} & c_{22} & \ldots & c_{m2} \\ \ldots & \ldots & \ldots & \ldots \\ c_{1n} & c_{2n} & \ldots & c_{mn} \end{pmatrix} \tag{10}$$

to yield this very compact form:

$$\mathbf{I} = \hat{\mathbf{s}}[\mathbf{C}]_{\hat{\mathbf{s}}} \tag{11}$$

with $\mathbf{I} \equiv (\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_m)$, $\hat{\mathbf{s}} \equiv \left(\hat{\mathbf{s}}_1, \hat{\mathbf{s}}_2, \ldots, \hat{\mathbf{s}}_n\right)$, and $[\mathbf{C}]_{\hat{\mathbf{s}}}$ is the large matrix of coefficients. This equation represents, in a single expression, the m spectra obtained by mixing the n solutions with concentrations c_{ij} for the i-th spectrum and the j-th solution.

1.7. Final notes on intensity spectra as vectors

If we have several components (i.e. methanol, ethanol, etc...) and there is no overlap whatsoever between the spectra (i.e the peaks are all distinct), then the base vectors $\hat{\mathbf{s}}_i$ and $\hat{\mathbf{s}}_j$ are **orthogonal**: their dot product is zero. However, it is more likely that the solutions *do* have overlapping spectra, therefore the base vectors (and consequently the base itself) will *not be orthogonal*. It is perfectly acceptable to have a base that is not orthogonal: it remains a base because any linear combination can create any spectrum we would measure.

1.8. Base change

The general expression for a vector as a function of its basis and its components in that basis is such that obviously, it stands correct for any basis:

$$\mathbf{I} = \mathbf{e}[\mathbf{C}]_{\mathbf{e}} = \mathbf{e}'\left[\mathbf{C}'\right]_{\mathbf{e}'} \tag{12}$$

It is the purpose of the present section to show how to go from a basis b to a basis b′, that is, how to transform the coefficients c into coefficients c′. For more information, you can look at the Youtube Video on base changes.

Since we can express any vector in a basis, we can choose to express the basis vectors $\mathbf{e}$ in the $\mathbf{e}'$ basis, with the coefficients $[\mathbf{Q}]_{\mathbf{e}'}$we do not know yet:

$$\mathbf{e} = \mathbf{e}'[\mathbf{Q}]_{\mathbf{e}'}, \tag{13}$$

where each column of the matrix $[\mathbf{Q}]_{\mathbf{e}'}$ is the component of the vector $\mathbf{e}_i$ in the $\mathbf{e}'$ basis. By definition, a basis set has enough vectors to cover the vector space, therefore both basis sets must have the same number of vectors, and the matrix $[\mathbf{Q}]_{\mathbf{e}'}$ is necessarily square, and can be inverted. We can therefore use (12) in (13) and obtain

simply:

$$\mathbf{I} = \mathbf{e}[\mathbf{C}]_{\mathbf{e}} = \left(\mathbf{e}'[\mathbf{Q}]_{\mathbf{e}'}\right)[\mathbf{C}]_{\mathbf{e}} = \mathbf{e}'\left([\mathbf{Q}]_{\mathbf{e}'}[\mathbf{C}]_{\mathbf{e}}\right) = \mathbf{e}'\left[\mathbf{C}'\right]_{\mathbf{e}'} \tag{14}$$

This means that, when the vectors in different bases are expressed by (12), the coordinates in the basis $\mathbf{e}'$ can be obtained from the components in the basis $\mathbf{e}$ by this simple transformation:

$$\left[\mathbf{C}'\right]_{\mathbf{e}'} \equiv [\mathbf{Q}]_{\hat{\mathbf{e}}'}[\mathbf{C}]_{\mathbf{e}} \tag{15}$$

1.9. Principal Component Analysis (PCA) in `sklearn`

The goal of Principal Component Analysis (PCA) is to obtain an orthogonal basis for a much smaller subspace than the original (it is a *dimensionality reduction* technique). We will identify this **orthonormal** PCA base as $\left\{\hat{\mathbf{p}}\right\}$, known as the principal component basis, or just the principal components.

At this point, it is farily simple to describe the process without worrying about the details: PCA takes a large number of samples spectra, and will:

1. Find the *principal components*, or an orthonormal basis $\left\{\hat{\mathbf{p}}\right\}$ that explains the variance of the data the best with a value that expresses how important they are.
2. Given a spectrum $\mathbf{I}$, it can return (*fit*) it to the PCA components and give the coefficients $[\mathbf{I}]_{\hat{\mathbf{p}}}$ in the PCA basis $\left\{\hat{\mathbf{p}}\right\}$, with $\mathbf{I} = \hat{\mathbf{p}}[\mathbf{I}]_{\hat{\mathbf{p}}}$.

So, because the present document is about the matheatical formalism first and foremost and that I do not want to dive so much into the Python details, let us just say that the following code will give us the principal components, and all the coefficients for our spectra in that base:

```
from sklearn.decomposition import PCA
#[...]
pca = PCA(n_components=componentsToKeep)
pca.fit(dataSet)                                # find the principal components
# The principal components are available in the variable pca.components_
# They form an orthonormal basis set
pcaDataCoefficients = pca.transform(dataSet) # express our spectra in the PCA
basis
# pcaDataCoefficients are the coefficients for each spectrum in the PCA basis
# Note: it is (c1*PC1+c2*PC2+....) + meanSpectrum = Spectrum
# as in equation (16) below.
```

1.10. PCA base change in `sklearn`

Equation (12) is not the only possibility to express a vector in different basis. We will see later that PCA often *translates* the sample vectors (i.e. the intensity spectra) to the "origin" by subtracting the mean spectrum from all spectra. This means that we have a more general transformation than (12) in that we do not express $\mathbf{I}$ in a different set of coordinates but rather $\mathbf{I} - \bar{\mathbf{I}}$:

$$\mathbf{I} - \bar{\mathbf{I}} = \hat{\mathbf{p}}[\mathbf{C}]_{\hat{\mathbf{p}}}, \tag{16}$$

with

$$\bar{\mathbf{I}} = \frac{1}{m}\sum_{j=1}^{m}\mathbf{I}_i \equiv \left(\hat{\nu}_1, \hat{\nu}_2, \ldots, \hat{\nu}_n\right)\left(\bar{\mathbf{I}}_1, \bar{\mathbf{I}}_2, \ldots, \bar{\mathbf{I}}_n\right)^T \equiv \hat{\nu}\left[\bar{\mathbf{I}}\right]_\nu \tag{17}$$

This average is computed in the "intensity" basis (or original basis $\left\{\hat{\nu}\right\}$) because that is the only basis we know when we start (i.e. we average all spectra). This small change where we subtract the mean is important, because to return to another basis used to generate $\mathbf{I}$, we need to write:

$$\mathbf{I} = \hat{\mathbf{p}}[\mathbf{C}]_{\hat{\mathbf{p}}} + \bar{\mathbf{I}} = \mathbf{e}[\mathbf{C}]_{\mathbf{e}}, \tag{18}$$

and if we try to follow the same development as in (14), we would quickly get stuck because we do not have $\bar{\mathbf{I}}$ neither in $\{\mathbf{e}\}$ or $\left\{\hat{\mathbf{p}}\right\}$ coordinates, we have it in the $\left\{\hat{\nu}\right\}$ coordinates :

$$\mathbf{I} = \hat{\mathbf{p}}[\mathbf{C}]_{\hat{\mathbf{p}}} + \hat{\nu}\left[\bar{\mathbf{I}}\right]_\nu = \mathbf{e}[\mathbf{C}]_{\mathbf{e}}, \tag{19}$$

Yet, this is the situation we will encounter later:

1. $\mathbf{I}$ is the many spectra we have acquired in the lab. They are in $\left\{\hat{\nu}\right\}$ basis (i.e. simple intensity spectra).
2. We can compute $\bar{\mathbf{I}}$ with (17), also in the spectral component basis $\left\{\hat{\nu}\right\}$.
3. The $\{\mathbf{p}_i\}$ basis is the Principal Component Analysis (PCA) basis that will be obtained from the module together with the coefficients $[\mathbf{C}]_{\hat{\mathbf{p}}}$. It comes from a singular value decomposition, and at this point, we do not worry oursleves with how it is obtained: we know we can obtain $\left\{\hat{\mathbf{p}}\right\}$ and $[\mathbf{C}]_{\hat{\mathbf{p}}}$ from `sklearn` and PCA.
4. Finally, the $\{\mathbf{e}_i\}$ basis will be our "solution" basis (or the *physically meaningful* basis) for which we would like to get the concentrations $[\mathbf{C}]_{\mathbf{e}}$ for our lab measurements. In Raman, this could be the lipid spectrum, DNA spectrum, protein spectrum etc... We know *some* $\{\mathbf{e}_i\}$ (from insight), but we certainly do not know them all. We want the coefficients to try to determine the concentrations of these molecules and get insight (or answers) from our experimental spectra, but we may not have all the components (i.e. we may

not have the full basis set).

There is mathematically not much we can do with these three coordinate systems in (19), unless we express the average spectrum $\bar{\mathbf{I}}$ in one or the other bases. We can do two things:

1. Express $\bar{\mathbf{I}}$ in the base $\{\mathbf{e}\}$
2. Express $\bar{\mathbf{I}}$ in the PCA base $\{\hat{\mathbf{p}}\}$

For reasons that should become clear later, we will choose to express $\bar{\mathbf{I}}$ in the base $\{\hat{\mathbf{p}}\}$ because, in fact, we do not know $\{\mathbf{e}\}$ completely, we only know *part* of it. If we knew $\hat{\mathbf{p}}\left[\bar{\mathbf{I}}\right]_{\hat{\mathbf{p}}}$, we could write:

$$\hat{\mathbf{p}}[\mathbf{C}]_{\hat{\mathbf{p}}} + \hat{\mathbf{p}}\left[\bar{\mathbf{I}}\right]_{\hat{\mathbf{p}}} = \mathbf{e}[\mathbf{C}]_{\mathbf{e}}, \tag{20}$$

$$\hat{\mathbf{p}}\left([\mathbf{C}]_{\hat{\mathbf{p}}} + \left[\bar{\mathbf{I}}\right]_{\hat{\mathbf{p}}}\right) = \mathbf{e}[\mathbf{C}]_{\mathbf{e}}, \tag{21}$$

If we define for clarity:

$$[\mathbf{C}_+]_{\hat{\mathbf{p}}} \equiv [\mathbf{C}]_{\hat{\mathbf{p}}} + \left[\bar{\mathbf{I}}\right]_{\hat{\mathbf{p}}}, \tag{22}$$

we can write:

$$\hat{\mathbf{p}}[\mathbf{C}_+]_{\hat{\mathbf{p}}} = \mathbf{e}[\mathbf{C}]_{\mathbf{e}}. \tag{23}$$

We obtain it by transforming the null spectrum $\mathbf{0}$ in equation :

$$\mathbf{0} = \hat{\mathbf{p}}[\mathbf{C_0}]_{\hat{\mathbf{p}}} + \bar{\mathbf{I}} \tag{24}$$

$$\bar{\mathbf{I}} = -\hat{\mathbf{p}}[\mathbf{C_0}]_{\hat{\mathbf{p}}} = \hat{\mathbf{p}}\left(-[\mathbf{C_0}]_{\hat{\mathbf{p}}}\right) \equiv \hat{\mathbf{p}}\left[\bar{\mathbf{I}}\right]_{\hat{\mathbf{p}}} \tag{25}$$